Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reality-anchored generative gating enables robust morphological classification and metacognitive triage of leukemia

Zhigang Yang

zhgyang@szu.edu.cn

Shenzhen University https://orcid.org/0000-0002-6530-4617

Shah Faisal

Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Zhao Hui

Department of Hematology, South China Hospital, Medical school, Shenzhen University, Shenzhen, 518116, P. R. China. No.1, Fuxin Road, Longgang District, Shenzhen, P. R. China, 518116

Zhangyue Wu

Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Zhang Shiqing

Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Liu Jia

Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Zou Yi

Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Yi Ye

Hematology Reagent R&D department, Mindray Bio-medical Electronics Co. LTD, Shenzhen, 518107, China

Zhang Ziyan

Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Huang Zhihui
Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Yang Kaiwen
Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Wu Meishi
Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Ye Qingyun
Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Ye Wenxuan
Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Wang Yizhe
Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Abdulkadir Abdulkadir
Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Lian Jiarong
Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Chen Gengwen
Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

Article

Keywords:

Posted Date: September 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10825472/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Reality-anchored generative gating enables robust morphological classification and metacognitive triage of leukemia

*Shah Faisal[1, #], Hui Zhao[2, #], Zhangyue Wu[1], Shiqing Zhang[1], Jia Liu[4], Yi Zou[1], Yi Ye[3], Ziyan Zhang[1], Zhihui Huang[1], Kaiwen Yang[1], Meishi Wu[1], Qingyun Ye[1], Wenxuan Ye[1], Yizhe Wang[1], Abdulkadir Zaheer Abdulkadir[1], Jiarong Lian[1], Gengwen Chen[3, *], and Zhigang Yang[1, *]*

[1] Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, China

[2] Department of Hematology, South China Hospital, Medical school, Shenzhen University, Shenzhen, 518116, China

[3]Hematology Product Development Department, Shenzhen Mindray Bio-Medical Electronics Co. Ltd., Shenzhen, Guangdong 518055, China

[4]Department of Laboratory Medicine, Jiangxi Maternal and Child Health Hospital, Nanchang, 330108, China

[#]These authors contributed equally to this work.

Corresponding Email: chengengwen@mindray.com; zhgyang@szu.edu.cn

Abstract

The morphological assessment of peripheral blood smears constitutes an indispensable cornerstone of leukemia subtyping and oncological triage. However, this critical diagnostic task remains labor-intensive, and the extreme intraclass heterogeneity of rare leukemic blasts persistently obstructs standard automated approaches. Conventional deep learning architectures struggle to bridge the gap between artificially balanced training distributions and raw clinical realities, often overfitting to the artifacts of generative data augmentation while failing to generalize to rare minority lineages. Here, we introduce a multi-objective generative-diagnostic framework that resolves these systemic vulnerabilities through reality-anchored feature gating. By synergizing a conditionally adapted latent diffusion model with a novel Generative Uncertainty-aware Retrieval Gate (GUR-Gate), our approach leverages generative morphological variation while utilizing strict angular margins to effectively deter non-physiological synthetic hallucinations. Our architecture decisively breaks the operational constraints of state-of-the-art discriminative models, achieving 99.30% overall accuracy, 97.57% macro F1-score, and 97.51% macro sensitivity across 11 distinct leukocyte sub-lineages in an independent clinical cohort. Furthermore, we embed heteroscedastic uncertainty calibration directly into the diagnostic pipeline, enabling the framework to autonomously flag highly ambiguous or anomalous blasts for expert human review. By decoupling generative expansion from explicit, reality-constrained representation learning, our framework provides a robust, self-calibrating, and interpretable paradigm for the automated differential diagnosis of leukemia.

Introduction

Leukemia and associated hematological malignancies constitute a severe global oncological burden, accounting for over 474,000 incident cases and approximately 311,000 disease-related

deaths annually worldwide,[1] and presenting a complex diagnostic challenge due to their extensive inter- and intracellular morphological variability. Standard automated cell counters provide efficient preliminary screening, definitive lineage subtyping, and the identification of rare pathological phenotypes, which require the high-resolution morphological evaluation of Wright-Giemsa-stained peripheral blood and bone marrow smears.[2] While the microscopic assessment serves as the established clinical gold standard, it is inherently labor-intensive, susceptible to cognitive fatigue, and constrained by significant inter-observer variability even among subspecialized hematopathologists. Furthermore, the sustainability of this manual workflow is currently challenged by shifting global workforce demographics. A widespread reduction in active, subspecialized hematopathologists, coupled with persistent recruitment deficits, has resulted in severely limited diagnostic capacity.[3] This deficit is most particularly evident in low- and middle-income countries (LMICs), where ratios frequently fall below one pathologist per 500,000 individuals. The intersection of an increasing global oncological incidence with a contracting diagnostic workforce routinely causes critical delays in treatment initiation.[4] Consequently, there is an essential clinical requirement to develop robust, automated computational diagnostic systems capable of providing expert-level morphological analysis at scale.

Contemporary hematological workflows heavily rely on automated multi-parameter flow cytometry and digital cell counters for high-throughput initial screening; however, these techniques inherently lack the high-resolution spatial and textural context required for fine-grained subtyping, routinely necessitating reflex manual microscopy for the definitive morphological classification of atypical leukemic blasts. To address the severe diagnostic constraints of manual smear review, deep convolutional neural networks (CNNs) have been increasingly deployed, demonstrating high accuracy in the automated classification of standard peripheral blood leukocytes.[5] Nevertheless, the clinical application of these supervised architectures is fundamentally inhibited by the inherent "long-tail" distribution of hematological malignancies. This difficulty is further compounded by the natural morphological heterogeneity across diverse patient demographics; real-world clinical presentations span extensive age ranges (e.g., 04 to 80 years) and include morphologically ambiguous pathological subtypes, such as Chronic Myelomonocytic Leukemia (CMML) and Acute Myeloid Leukemia (AML-M5).[6] Traditional data augmentation strategies and early-generation Generative Adversarial Networks (GANs) have proven inefficiency in addressing this severe class imbalance, as they frequently induce mode collapse, distort crucial cytological textures, or generate clinically invalid morphological artifacts. Although advanced Latent Diffusion Models are highly effective at preserving cellular structures, they still introduce a slight domain shift when compared to authentic blood smears.[7] Unless these generated features are explicitly tied to real clinical ground truths during training, the classifier may end up memorizing artificial boundaries rather than true diagnostic markers. Furthermore, conventional CNNs operate as uninterpretable algorithmic "black boxes" that optimize strictly for categorical cross-entropy, failing to ground their predictions in quantitative biological metrics or provide self-calibrated uncertainty estimations when evaluating degraded or morphologically ambiguous smears.[8] Consequently, there is a critical necessity to transition beyond standard classification paradigms toward integrated, multi-objective frameworks that concurrently leverage advanced generative synthesis for biologically realistic class parity and uncertainty-weighted architectural designs to ensure diagnostic interpretability and clinical safety.

The morphological characterization of hematological malignancies is intrinsically open-ended and frequently lacks a definitive visual ground truth.[9] Microscopic appearances can be highly anomalous, with critical clinical implications hidden within subtle textural or structural variations. Furthermore, pathological phenotypes are often obscured by irrelevant biological artifacts or instrumental variations originating from disparate staining protocols and optical equipment.[10] Consequently, manual classification is a challenge that human experts can only imperfectly meet, inevitably exhibiting marked inter-observer variability. While recent work has successfully applied standard CNN to automate blood cell assessment, such as the classification of leukemia lineages and the differentiation of bone marrow cells,[11] these purely discriminative approaches have significant translational limitations. Standard CNN primarily learns mathematical decision boundaries optimized for categorical cross-entropy. They are not inherently designed to capture the true, underlying biological distribution of cellular appearances,[12] nor are they equipped to reason about the physical metrics of the cells they analyze.[8]

To safely automate hematological diagnostic systems, a desirable clinical architecture must address specific operational constraints beyond standard categorical accuracy. A clinically viable model must first demonstrate robustness against severe data imbalance, maintaining high diagnostic sensitivity for rare pathological variants without degrading overall performance. Furthermore, to ensure essential clinical interpretability, the network's predictive manifold should be explicitly grounded in verifiable cytological features, such as nuclear contour, cytoplasmic granularity, and chromatin density, rather than in opaque latent representations.[13] Finally, diagnostic safety mandates the integration of probabilistic self-calibration. Rather than relying on deterministic softmax confidence scores, the architecture must quantitatively estimate its own predictive uncertainty.[14] This mathematical safeguard ensures the systematic identification of morphologically ambiguous or out-of-distribution cells, reliably deferring them for expert hematological review and thereby mitigating the risks associated with automated misclassification.[8] However, current monolithic CNN architectures struggle to satisfy these criteria simultaneously. They are highly susceptible to the "long-tail" distribution of clinical data, overfit rapidly on minority classes, and operate as uninterpretable black boxes that output overconfident, uncalibrated softmax probabilities.[15]

To address the dual challenges of data scarcity and diagnostic opacity in clinical hematology, we introduce an advanced Generative-Diagnostic framework that explicitly decouples the modeling of morphological distribution from the execution of diagnostic reasoning. First, to resolve the limitations of data efficiency and robustness, we employ a Conditional Latent Diffusion Model (CLDM)[16] enhanced with Low-Rank Adaptation (LoRA)[17] and Textual Inversion. [18] Instead of merely applying geometric augmentations, the CLDM acts as a highly constrained generative engine, synthesizing the complete morphological distribution of rare cell types to achieve perfect class parity. Second, we perform classification using a specialized Diagnostic ResNet-50[19] trained on this balanced dataset. Because synthetic images can still differ slightly from real patient samples, we designed a novel Margin-Penalized Retrieval Feature Gate to bridge this gap. During training, this mechanism pairs each generated cell with its closest real-world clinical reference. It then aligns them using an ArcFace-style angular similarity[20] with a stable margin penalty. By actively anchoring the synthetic features to authentic clinical data, this gate prevents the network from learning synthetic artifacts. Moving beyond simple boundary discrimination, this network is architected with a multi-objective head

that simultaneously classifies cell lineage, regresses quantitative morphological metrics, maps cells into a contrastive hypersphere, and computes predictive log-variance ($S = log\sigma^2$). The proposed modelling approach, and the specific contributions of this work are: (1) the novel application of a CLDM framework to resolve extreme class imbalance in a demographically diverse (ages 04–80), real-world hematological data, synthesizing a mathematically balanced, expert-validated cohort of 15,881 images; (2) the introduction of a Margin-Penalized Retrieval Feature Gate, which uses stable ArcFace-style angular margins to actively tie synthetic cellular features to authentic clinical references, preventing the network from learning visual artifacts, (3) the design of a multi-objective diagnostic architecture that leverages multi-level feature fusion to simultaneously optimize for classification and metric learning; (4) the integration of an auxiliary morphological regression head that actively penalizes the network for learning visual features lacking physical biological validity; (5) the implementation of a heteroscedastic uncertainty calibration framework, providing a numerical safety guardrail that identifies out-of-distribution or ambiguous cells for human review; and (6) a comprehensive evaluation framework deployed across 11 distinct normal and malignant leukocyte classes, demonstrating near-perfect diagnostic discrimination (99.30% accuracy) to bridge the gap between raw computational accuracy and clinical diagnostic trustworthiness.

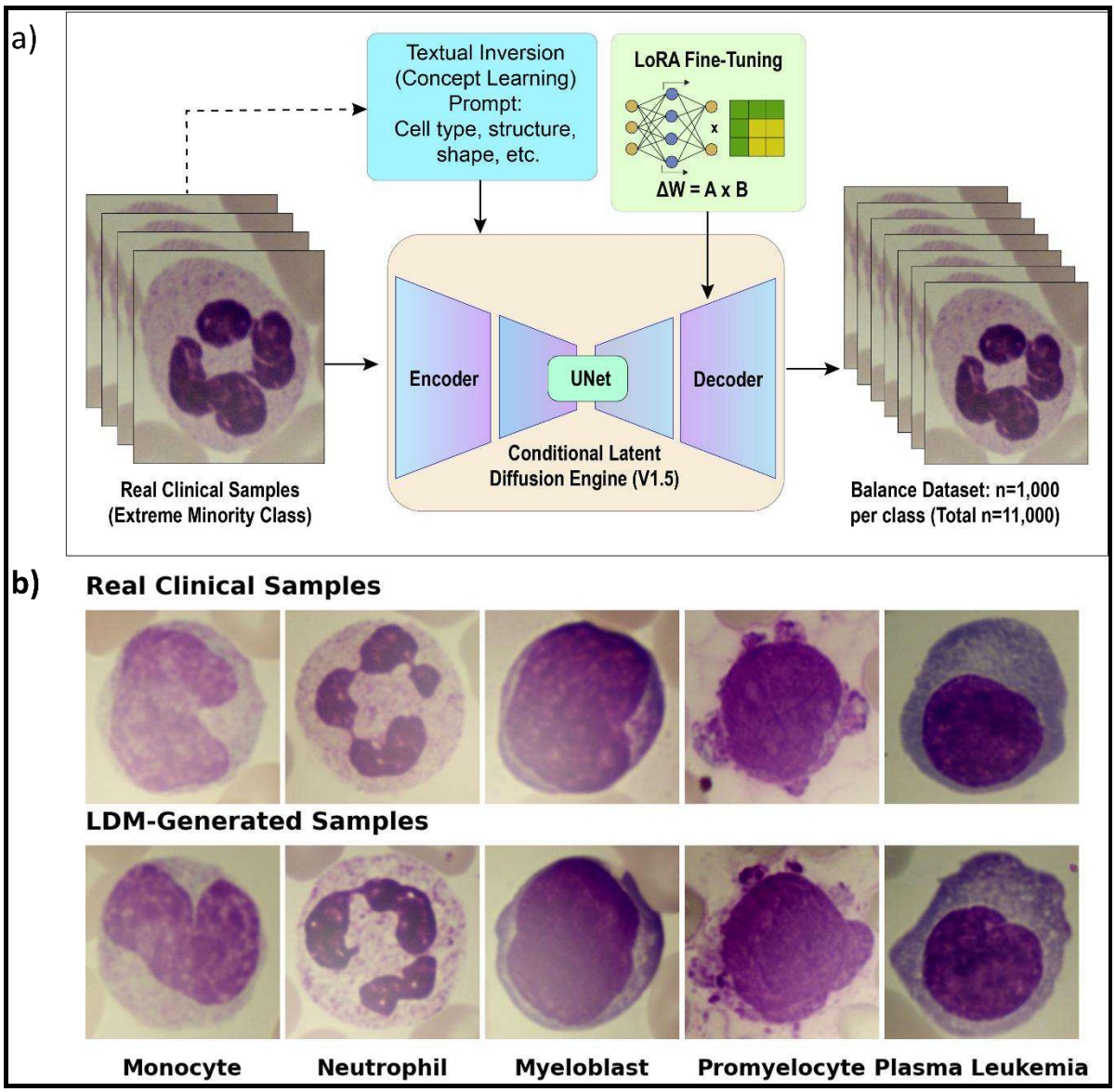


Figure 1 | Architectural overview of the LoRA-enhanced Conditional Latent Diffusion Model (CLDM) for morphological synthesis. **a,** Schematic representation of the forward diffusion and latent encoding processes, detailing the integration of Textual Inversion and Low-Rank Adaptation (LoRA). The cross-attention mechanism within the reverse denoising U-Net is utilized to structurally condition specific leukocyte sub-lineages and morphological phenotypes. **b,** Qualitative demonstration of the structurally conditioned image-to-image (img2img) translation strategy, enabling the synthesis of a mathematically balanced, expert-validated clinical manifold (n = 15,881) from an inherently imbalanced baseline cohort.

Methods and Materials

Chromatic-Based Localization and Adaptive Dataset Standardization

To bridge the gap between high-resolution peripheral blood smear (PBS) scans and the localized requirements of deep generative models, we developed a specialized pre-processing pipeline centered on Chromatic-Based Localization (CBL). The fundamental challenge in pathological image analysis is the presence of confounding background elements, such as overlapping erythrocytes and plasma artifacts. We addressed this by leveraging the biochemical properties of standard hematological stains (e.g., Wright-Giemsa), which exhibit a high affinity

for the acidic components of the leukocyte nucleus and cytoplasm, resulting in a distinct spectral signature in the purple-to-magenta range.

We map the input image $\boldsymbol{I_{raw}}$ into the Hue-Saturation-Value (HSV**)** color space to decouple intensity from chromaticity. A binary mask M is generated using a selective thresholding operation:

$$M(x,y) = \begin{cases} 1 \; if \; H_{lower} \leq H(x,y) \leq H_{upper} \\ 0 \; \; otherwise \end{cases}$$

where the hue range $[H_{lower}, H_{upper}]$is tuned to the characteristic absorption peak of stained chromatin (typically 120° to 160° in standard color mapping). To ensure the inclusion of delicate cytoplasmic structures such as pseudopodia or the thin cytoplasmic rim often found in leukemic blasts, the mask underwent morphological refinement using a 5x5 structuring element for dilation and closing operations.[21] Finally, an adaptive bounding box was extracted with a fixed padding of 10 pixels. This process standardized the dataset into semantically dense, cell-centric crops, ensuring that subsequent latent space operations were focused exclusively on diagnostically relevant morphological features

Morpho-Semantic Synthesis via LoRA-Enhanced Latent Diffusion

The core of our data augmentation strategy is a CLDM architecture, chosen for its ability to model complex distributions in a compressed representation space.[16] While standard generative models often produce visually plausible images that lack clinical validity, our framework ensures morphological accuracy through a dual-optimization approach involving Textual Inversion (TI) and LoRA. By operating in the latent space Z, where z = ε(x) is the representation encoded by a pre-trained autoencoder, the model focuses on high-level cytological structures rather than pixel-level noise, significantly enhancing the fidelity of synthesized leukemic morphologies.

To encode rare pathological features that are absent from general linguistic models (e.g., Auer rods in AML-M3 or nuclear clefting in certain lymphomas), we implemented Textual Inversion.[18] This process introduces a learnable embedding vector v^* associated with a unique pseudo-word token S^* for each of the 11 cell classes. During training, the U-Net and autoencoder weights remained frozen while v^* was optimized to minimize the latent diffusion loss:

$$\boldsymbol{L_{LDM}} = \boldsymbol{E}_{\boldsymbol{z}\sim\boldsymbol{\varepsilon}(\mathbf{x}),\mathbf{y},\in\sim\mathbf{N}(\mathbf{0},\mathbf{1}),\mathbf{t}}[\|\in-\in_{\boldsymbol{\theta}}(\boldsymbol{z_t},\boldsymbol{t},\boldsymbol{c_\psi}(\boldsymbol{S}^*))\|_2^2]$$

where c_ψ represents the text encoder mapping the learned token to a conditioning vector. To further refine the model's response to these learned tokens and preserve the specific textural and staining characteristics of our clinical data, we applied LoRA to the cross-attention layers of the U-Net.[17] By injecting rank-decomposed matrices into the query and value projections, the model learned to modulate its attention based on the specific cell type being synthesized. This combination allowed for the generation of a balanced, high-fidelity synthetic dataset where rare leukemic classes achieved parity with normal cell populations, providing a robust foundation for the training of the subsequent classification architectures.

Multi-Level Feature Fusion Architecture

Standard convolutional neural networks, which typically perform Global Average Pooling (GAP) solely on the final layer,[22] are suboptimal for hematological analysis. While the deepest layers effectively capture high-level semantic shapes (e.g., overall nuclear contour), they routinely discard the fine-grained, low-level textural details (e.g., cytoplasmic granularity and chromatin stippling) that are critical for differentiating ambiguous leukemic blasts.[23]

To resolve this, our diagnostic engine is constructed upon a specialized ResNet-50 backbone designed for Multi-Level Feature Fusion.[24] Rather than relying on a single output layer, we systematically extract continuous feature maps from three distinct depths of the network: Layer 2 (F_{L2} for granular texture), Layer 3 (F_{L3} for intermediate structures), and Layer 4 (F_{L4} for comprehensive cellular morphology). Following adaptive pooling, these multi-scale features are concatenated into a unified tensor and routed through a shared fusion block comprising a linear transformation, 1D Batch Normalization, a ReLU activation, and Dropout (**Figure 2a**). This process yields a densely integrated, 1024-dimensional bottleneck representation ($Z_{fusion} \in \mathbb{R}^{1024}$). By explicitly forcing the network to simultaneously evaluate both micro-textures and macro-shapes, this unified workspace provides a significantly richer foundation for the downstream multi-task diagnostic heads. However, a naive concatenation assigns equal static weight to all feature channels, which is suboptimal given that the diagnostic importance of specific textures or shapes varies drastically across different leukemic cell classes. To explicitly model the interdependencies between these multi-scale features, we integrated a Channel Attention Gate immediately following the fusion block. This mechanism utilizes a Squeeze-and-Excitation architecture to dynamically recalibrate the fused representation. The 1024-dimensional vector is first compressed through a linear bottleneck to 64 dimensions (utilizing a reduction ratio of 16) and passed through a ReLU activation. A subsequent linear layer restores the dimensionality to 1024, followed by a Sigmoid activation to generate a continuous attention mask. This mask is multiplied element-wise against the original fused vector. Consequently, the network learns to adaptively amplify highly discriminative diagnostic channels—such as those encoding pathognomonic nucleoli or granular anomalies—while actively suppressing irrelevant domain noise prior to routing the refined feature vector to the retrieval-guided gate and final classification heads.

Margin-Penalized Retrieval Feature Gate

While the CLDM successfully resolves numerical class imbalance, synthetic images are inherently susceptible to subtle domain shifts. If a network optimizes directly on these generated samples without clinical constraint, it risks memorizing synthetic visual artifacts rather than genuine biological markers. To counteract this, we introduce a novel Margin-Penalized Retrieval Feature Gate that actively anchors synthetic representations to authentic clinical ground truths during the forward pass.

For every synthetic feature vector (z_{syn}) passing through the fusion block, the gate dynamically retrieves its nearest real-world clinical reference (r_{real}) from the training manifold. To enforce strict alignment between the synthetic and real domains, we compute an ArcFace-style angular similarity. However, standard angular margin implementations that rely on the inverse cosine function ($ArcCos$) frequently suffer from gradient explosions and NaN (Not a Number) errors during mixed-precision training when similarities approach ± 1.[25]

To ensure absolute mathematical stability, we designed a computationally safe trigonometric

margin penalty utilizing the Pythagorean identity. First, both vectors are L_2-normalized, and their standard cosine similarity is computed and safely clamped:

$$Cos\theta = \frac{z_{syn} . z_{real}}{\|z_{syn}\|_2 \|z_{real}\|_2}$$

Instead of extracting the angle directly, the corresponding sine is derived:

$$Sin(\theta) = \sqrt{\max(0, 1 - cos^2(\theta))}$$

A strict angular margin (m = 0.5) is then applied using the stable angle addition formula to calculate the penalized similarity:

$$\cos(\theta + m) = \cos(\theta)\cos(m) - \sin(\theta)\sin(m)$$

This penalized similarity score ($S_{penalized}$), along with the original vectors (z_{syn} and z_{real}) and their absolute difference ($|z_{syn} . z_{real}|$) is concatenated and passed through a multi-layer perceptron (MLP) with a Sigmoid activation to predict a continuous, soft gating mask ($g \in [0,1]$). The final, artifact-suppressed representation is thus calculated as:

$$z_{final} = g \odot z_{syn} + (1 - g) \odot r_{real}$$

As detailed in **Figure 2b**, this mathematically stable gating mechanism explicitly pulls generative distortions back toward verifiable clinical realities in the latent space, drastically improving the network's generalizability on the isolated testing cohort.

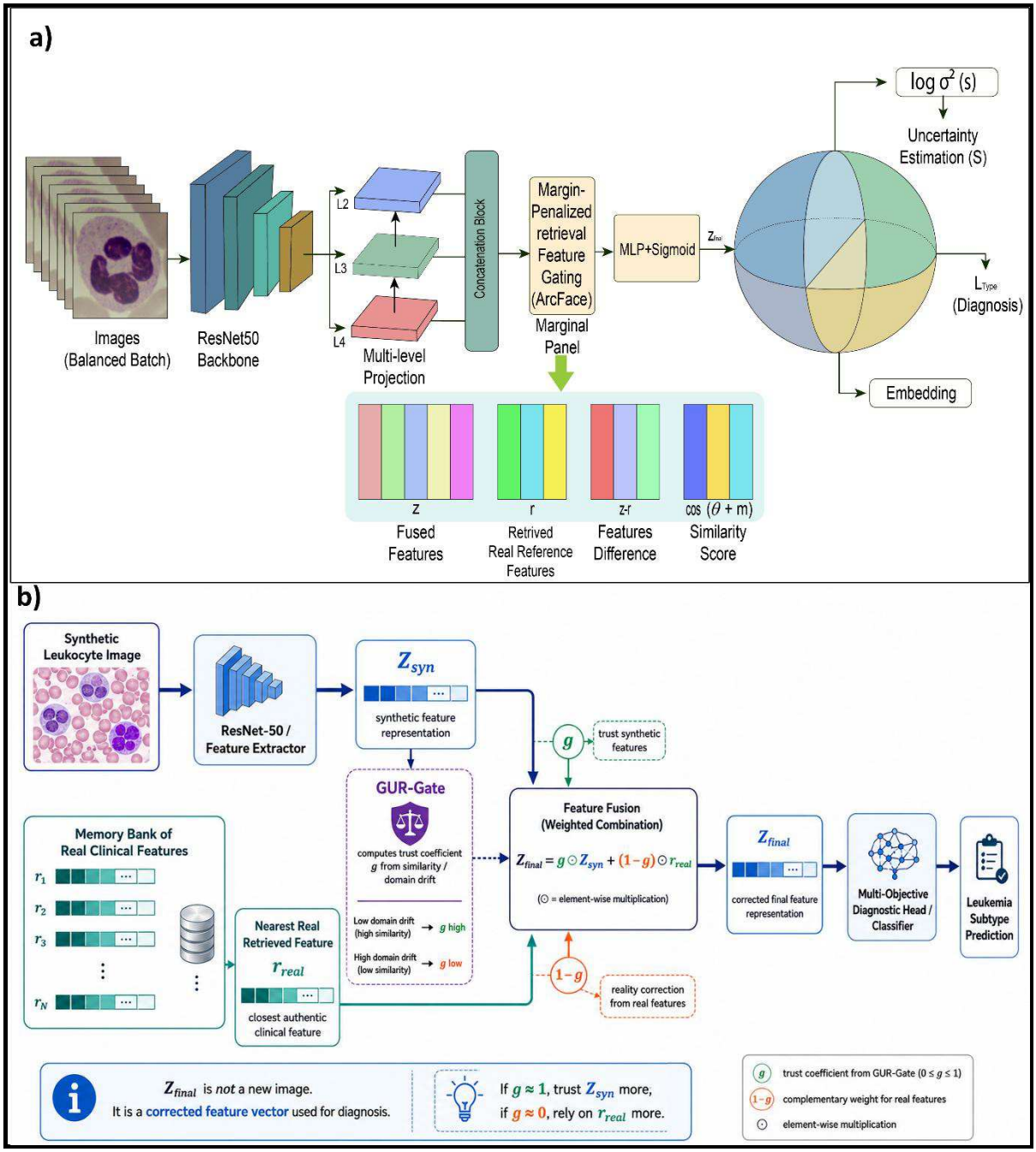


Figure 2 | Architectural overview of the Multi-Objective Diagnostic Framework and Generative Uncertainty-aware Retrieval Gate (GUR-Gate). **a,** Macro-architecture of the multi-level feature fusion diagnostic engine. Continuous feature maps are systematically extracted from distinct network depths (Layers 2, 3, and 4) of a ResNet-50 backbone to simultaneously capture low-level cytoplasmic granularity and high-level holistic cellular morphology. The concatenated features pass through the GUR-Gate to yield an artifact-suppressed bottleneck representation. This shared latent workspace is then routed through parallel multi-objective heads optimizing for categorical diagnosis, contrastive hyperspherical embedding, continuous morphological regression, and aleatoric uncertainty estimation ($S = log\sigma^2$. **b,** Detailed mechanics of the Margin-Penalized Retrieval Feature Gate (GUR-Gate). To explicitly suppress generative hallucinations and domain shifts, the gate dynamically retrieves the nearest authentic clinical ground truth (r_{real}) from a memory bank for every synthetic feature vector (z_{syn}). An ArcFace-

style angular similarity and stable trigonometric margin penalty are computed to predict a continuous trust coefficient (g). The final representation z_{final} is an element-wise, reality-corrected feature vector, actively anchoring the synthetic manifold to verifiable clinical biology prior to downstream diagnostic reasoning.

Multi-Task Diagnostic Classification Architecture

The diagnostic engine executes multi-objective predictions via four parallel inferential branches. The artifact-suppressed representation ($z_{final} \in \mathbb{R}^{1024}$) Produced by the feature gate serves as a shared latent workspace for these concurrent tasks (**Figure 2a**). The primary classification branch utilizes a linear transformation to map the latent features into 11 distinct categorical logits, computing the final diagnostic probabilities for the leukocyte sub-lineages. To explicitly ground these visual predictions in hematological reality, the architecture incorporates an auxiliary morphological regression head. This branch predicts a continuous vector of quantitative cytological metrics ($m \in \mathbb{R}^d$) such as nucleocytoplasmic ratio and cytoplasmic granularity, effectively penalizing the network for relying on visual features that do not correlate with established physical geometry.

Concurrently, the architecture integrates a metric-learning component through a contrastive embedding head. This branch projects the latent features into a lower-dimensional hypersphere ($e \in \mathbb{R}^{128}$), which is rigorously L_2-normalized.[26]

$$e = \frac{W_c z_{final} + b_c}{\left\| W_c z_{final} + b_c \right\|_2}$$

This normalization ensures that the similarity between cells is measured via angular distance on a unit manifold, facilitating the robust clustering of rare leukemic variants. Finally, an uncertainty calibration head processes the shared representation to compute a sample-specific predictive log-variance,[27] serving as a continuous internal confidence metric for each input image.

Integrated Diagnostic Loss Function

The optimization of the Diagnostic ResNet is governed by a multi-task loss function $\mathcal{L}_{\text{total}}$, which balances categorical classification, continuous morphological regression, and predictive uncertainty calibration. The primary classification loss utilizes a standard categorical cross-entropy ($\mathcal{L}_{CE}$) formulation to establish robust decision boundaries between the 11 leukocyte sub-lineages. However, to fulfill the clinical requirement for diagnostic safety, this baseline classification loss is dynamically weighted by the network's predictive stochastic uncertainty. For a given sample, the uncertainty calibration head outputs a continuous log-variance parameter ($S = log\sigma^2$) The uncertainty-weighted classification loss is thus formulated as a Gaussian likelihood.[27]

$$\mathcal{L}_{\text{class}} = \frac{\exp(-s)}{2} \mathcal{L}_{CE} + \frac{s}{2}$$

This mechanism forces the model to self-calibrate. If the network encounters a highly anomalous, morphologically ambiguous, or out-of-distribution cell, it led to yield a substantial predictive variance (s). This mathematically attenuates the cross-entropy penalty for that specific sample, preventing uncertain, noisy gradients from destabilizing the learned network

weights. Beyond categorical classification, the overall objective actively grounds the latent space in verifiable physical biology. We incorporate a Mean Squared Error (MSE) morphological regression penalty ($\mathcal{L}_{\text{morph}}$) to ensure the network accurately maps latent vectors to physical cytological traits (e.g., cytoplasmic granularity), alongside a metric-learning contrastive loss ($\mathcal{L}_{\text{cont}}$) to enforce tight intra-class clustering on the unit hypersphere. The final composite loss is a parameterized summation of these distinct objectives:

$$\mathcal{L}_{\text{total}} = \lambda_{type}\mathcal{L}_{\text{class}} + \lambda_{morph}\mathcal{L}_{\text{morph}} + \lambda_{cont}\mathcal{L}_{\text{contrastive}}$$

By simultaneously optimizing for robust boundary discrimination, physical trait regression, and uncertainty estimation, the architecture establishes a highly interpretable and self-calibrating clinical manifold.

Clinical Cohort and Data Acquisition

Clinical Cohort and Patient Demographics This prospective, observational study was conducted utilizing peripheral blood and bone marrow samples acquired at South China Hospital, Shenzhen. Ethical approval was granted by the Institutional Review Board of South China Hospital, and informed consent was obtained from all participants prior to data collection. To ensure the computational framework was evaluated against true clinical heterogeneity, the acquisition cohort encompassed a diverse patient demographic, comprising male and female patients with an age distribution ranging from 04 to 80 years.

The cohort captured a broad spectrum of hematological presentations. Baseline clinical ground truths were established through rigorous pathological examination by board-certified hematopathologists, verified by standard flow cytometry where clinically indicated. The patient population was categorized into individuals with normal baseline blood profiles, those with reactive conditions, and those presenting with rare, morphologically ambiguous pathological subtypes, such as Chronic Myelomonocytic Leukemia (CMML) and Acute Myeloid Leukemia (AML-M5).

Sample Processing and Image Acquisition Standardized peripheral blood smears were prepared and stained using a Wright-Giemsa protocol (Zhuhai Baisuo, China) to ensure distinct nuclear and cytoplasmic differentiation. Digital acquisition was performed using a Leica DM750 biological microscope under 1000× total magnification. To maintain longitudinal consistency for subsequent deep learning tasks, images were captured utilizing the Beion V4.90 medical imaging system under strictly standardized illumination and exposure parameters. The resulting dataset consisted of high-resolution PNG images characterized by high morphological fidelity and uniform staining. (**Supplementary Note 1 and Supplementary Figure S1**).

Clinical Manifold and Extreme Class Imbalance To streamline clinical taxonomy while maintaining diagnostic relevance, overlapping disease presentations (e.g., Multiple Myeloma and Plasma Cell Leukemia) were unified, resulting in 11 distinct normal and malignant leukocyte sub-lineages. The baseline clinical dataset demonstrated a pronounced "long-tail" distribution typical of real-world hematology.[28] Normal leukocyte populations were significantly prevalent, comprising Monocytes (n=135), Lymphocytes (n=120), Eosinophils (n=184), Basophils (n=129), and Neutrophils (n=189). Conversely, the pathological cohort exhibited extreme scarcity, particularly across acute myeloid leukemia (AML) subtypes: Myeloblasts (AML-M1/M2, n=41), Abnormal Promyelocytes (AML-M3, n=45),

Promonocytes (AML-M5, n=35), Monoblasts (AML-M5, n=50), Lymphoblasts (ALL, n=56), and the combined Plasma Cell/PCL category (n=344). This yielded a raw clinical manifold of 1,333 authentic microscopic images.

Layered Generative Expansion and Data Partitioning Strategy. To resolve this extreme class imbalance without inducing the overfitting associated with traditional oversampling, we utilized our LoRA-enhanced CLDM, as shown in Figure 1a, to synthesize a mathematically balanced manifold. A specialized, layered data partitioning strategy was implemented to rigorously preserve patient-level stratification and prevent data leakage during evaluation.[29] First, 618 pathological images were selected as an independent set of authentic clinical ground truths. To synthesize pathologically accurate variants, the LoRA-conditioned LDM was trained, ensuring no leakage of data. Images were generated for each minority class through an image-to-image (img2img) translation strategy, reaching classes parity (e.g., 15881 images for pathological categories) as shown in **Figure 1b**.

This synthesized data was then used to construct the final balanced "internal" sets:

- **Final Training Set:** Combined the 448 real minority images with corresponding generated samples.
- **Final Validation Set:** Combined the 56 real minority validation images with corresponding generated samples.
- **Final Test Set:** Combined the 114 real minority test images with corresponding generated samples.

External Validation Cohort and Expert Review Protocol The finalized framework was evaluated across 11 leukocyte classes using two distinct strategies. First, on the "internal" combined (real minority + generated) test set. Second, on an "external" pure original batch of images from the hospital, which had been entirely held out from all aspects of model training, including image generation.

To ensure diagnostic and morphological integrity, the expanded synthetic manifold was subjected to a rigorous, multi-stage expert review protocol. Certified hematologists conducted blinded clinical validations of the generated outputs. A strict exclusionary protocol was enforced: any synthetic sample demonstrating specific chromatin density distortion, absent cytoplasmic granularity, optical blurring, or biologically impossible physical geometries was entirely excluded from the final training manifold. This dual-source validation ensured that the 15,881 image cohort utilized to train the Diagnostic ResNet50 was not only numerically balanced but of clinical-grade fidelity.

Given the inherently limited size of the clinical fine-tuning dataset, it was imperative to enforce rigid structural constraints during synthesis to guarantee morphological stability. Consequently, we employed a structurally conditioned image-to-image (img2img) generative strategy. This approach allowed us to precisely control the phenotypic expression of the synthesized leukocytes for instance, directing the morphological transition between lymphocytic and neutrophilic lineages by systematically modulating the textual conditioning prompts **(Supplementary Table S1 and Supplementary Figure S2a, b.** Furthermore, to explicitly validate the necessity of our architectural approach, we conducted a comparative ablation. We evaluated images generated using the foundational, unadapted Stable Diffusion model against our framework using identical text prompts. As demonstrated in

Supplementary Figure S2c, the unadapted baseline model fails to capture specialized hematological textures, conclusively proving that LoRA fine-tuning is strictly required to successfully cross the domain gap into high-fidelity pathological generation.

Results and discussion

We begin our evaluation by establishing the clinical integrity of the expanded 15881 image leukocyte dataset, which was strictly partitioned into a 70:10:20 ratio for training, validation, and testing to prevent data leakage. To ensure the generative process did not introduce hallucinatory noise, the synthesized pathological variants were first qualitatively examined by hospital hematology experts to confirm the preservation of true biological morphological features rather than generative artifacts. Next, we quantitatively validate this structural fidelity through deep feature cosine similarity analysis, mathematically proving that the synthetic cohorts reside on the same manifold as the real clinical data. Following this data validation, we assess the in-domain diagnostic efficacy of the multi-objective ResNet50 across all 11 normal and malignant sub-lineages, specifically focusing on its ability to resolve long-tail class imbalance. We then investigate the network's capacity for morphological grounding by analyzing its contrastive latent space (t-SNE) and physical trait predictions. Subsequently, we examine the framework's heteroscedastic uncertainty calibration, demonstrating its metacognitive ability to numerically flag ambiguous predictions for human safety review.

3.1 Qualitative Validation of the Generative Cohort

To establish the morphological authenticity and clinical viability of the synthesized leukocyte variants, we conducted a blinded, two-phase qualitative evaluation with a panel of three independent expert hematologists. The evaluation was designed to test both the diagnostically recognizable features of the cells and their structural indistinguishability from real clinical samples. Crucially, during the entire evaluation process, the expert panel did not reject a single generated image, confirming the absence of structural distortions, optical blurring, or biologically impossible artifacts in the synthetic manifold.

Diagnostic Feature Preservation In the first phase, the experts were tasked with independently diagnosing the cellular lineage of 200 randomly selected images encompassing multiple leukocyte classes. The objective was to confirm that the generated features strictly adhered to known hematological criteria. Expert 1 achieved a diagnostic accuracy of 97.0% (194/200), while Expert 2 and Expert 3 correctly labeled 77.0% (154/200) and 79.5% (159/200) of the images, respectively, presented in (**Supplementary Table S2)**. This inter-observer variance is consistent with the established difficulties of manual microscopic review in complex hematological sub-lineages. However, the high baseline protocol confirms that the generative model successfully preserved the essential cytological markers required for clinical evaluation.

The Visual Turing Test. To explicitly measure the "domain gap" between the real and synthetic manifolds, the experts were subjected to a Visual Turing Test. They were presented with a separate, unlabeled batch of 220 images (110 authentic clinical images and 110 generated images) and asked to distinguish the synthetic cells from the real ones. The results demonstrated profound difficulty in visually separating the domains, indicating high synthetic fidelity. Expert 1 correctly identified only 10 of the 110 generated images and 25 of the 110 real images. Expert 2 similarly identified only 10 of the 110 generated images, while classifying 98 of the 110 real images correctly. Expert 3 exhibited higher sensitivity to generative artifacts,

identifying 87 of the 110 generated images and 55 of the 110 real images **(see Supplementary Table S2 for full expert concordance data)**. The extreme variability across the panel and specifically the failure of Experts 1 and 2 to flag 100 out of 110 synthetic images strongly suggests that the conditionally generated cells are visually and biologically identical to standard Wright-Giemsa-stained hospital smears, validating the synthetic cohort for subsequent diagnostic training.

3.2 Quantitative Validation of Structural Fidelity and Latent Alignment

To rigorously confirm that the visual fidelity observed by the expert hematology panel translates into structural and mathematical authenticity, we evaluated the generative manifold using a combination of qualitative morphological grids, t-SNE latent space projections, and deep feature cosine similarity.

First, qualitative structural comparisons between the authentic clinical samples and the Conditional Latent Diffusion Model (CLDM) generated cohorts demonstrate exceptional preservation of fine-grained cytological details, including nuclear contours, chromatin density, and cytoplasmic granularity across diverse sub-lineages **(Figure 3a)**.

To determine if these generated features successfully map to the correct biological distributions, we projected the high-dimensional feature embeddings into a 2D contrastive latent space via t-SNE **(Figure 3b)**. The projection reveals a highly structured topological alignment. Rather than forming isolated generative clusters or suffering from mode collapse, the synthetic leukocyte variants seamlessly integrate into the authentic clinical clusters. The 11 distinct lineages maintain strict inter-class separability, mirroring the biological taxonomy of the cells, while the synthetic samples perfectly populate the intra-class distributions of their real-world counterparts.

To quantify this visual manifold alignment, we computed the cross-domain deep feature cosine similarity **(Figure 3c)**. The analysis revealed exceptional structural correlation: the macro-average cross-domain similarity (Real-to-Synthetic) was 0.7978, tightly mirroring the authentic intra-class baseline (Real-to-Real mean: 0.8058), resulting in a negligible overarching domain gap of just 0.008. Crucially, for notoriously difficult minority classes such as Abnormal Promyelocytes (M3) and Promonocytes (AML-M5), the synthetic generated features exhibited marginally higher intra-class similarity (0.8056 and 0.7622, respectively) compared to the real-world clinical baselines (0.8021 and 0.7529). This phenomenon suggests that the conditional LDM successfully distilled and generated canonical cytological markers rather than memorizing individual clinical noise. Together, these metrics mathematically guarantee that the synthetic cohorts reside on the identical high-dimensional feature space as This phenomenon indicates that the conditional LDM effectively distilled and produced standard cytological markers. The clinical ground truths, validating the expanded dataset for downstream diagnostic optimization.

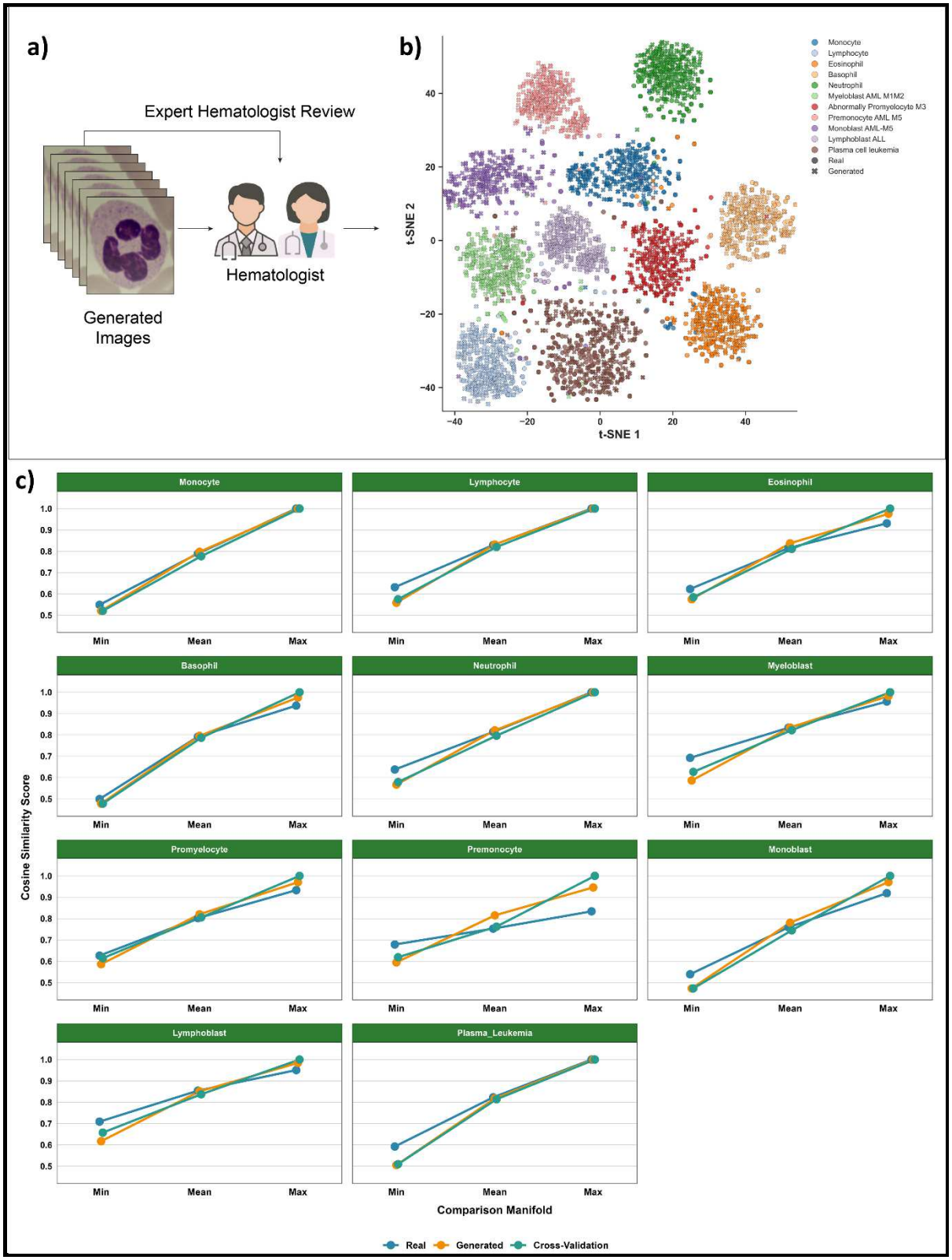


Figure 3 | Structural fidelity and topological alignment of the conditionally generated clinical manifold. **a,** High-resolution qualitative visual grids comparing authentic clinical samples with synthetic variants across diverse leukocyte sub-lineages, demonstrating the preservation of fine-grained cytological markers. **b,** A 2D t-SNE projection of the contrastive latent space, revealing seamless topological integration and strict inter-class separability between the 11 synthetic and real-world leukocyte lineages. **c,** Quantitative deep feature cosine similarity analysis. The synthetic cohorts achieve a macro-average

cross-domain similarity of 0.7978, mapping tightly to the authentic intra-class baseline (0.8058) and confirming a negligible domain gap.

3.3 Diagnostic classification and resolution of class imbalance

To evaluate the clinical viability and generalization capacity of the GUR-Gate ResNet architecture, diagnostic performance was assessed across two distinct testing regimes: an internal holdout set encompassing the expanded generative manifold and an independent external clinical cohort.

In-domain performance and class imbalance

The model was initially evaluated on a 20% stratified internal holdout set (n = 3174), comprising both authentic and conditionally generated images, to verify algorithmic convergence on the expanded feature space. The network achieved an overall accuracy of 99.56%, with macro-averaged precision of 99.59%, recall, and F1-scores converging at 99.56% (**Figure 4a**). Notably, this phase demonstrated the resolution of the dataset's long-tail class imbalance. The architecture established robust decision boundaries for rare pathological precursors, yielding a recall of almost 1.00 for myeloblast (AML-M1M2 n= 204), monoblasts (AML-M5, n = 435), promonocyte (AML-M5, n = 415), and promyelocytes (M3, n = 205). These results indicate that the ArcFace-style angular margin penalty effectively stabilized feature representation for minority classes without underfitting.

Diagnostic sensitivity was maintained for critical leukemic states, achieving a recall of over 0.99 to 1.00 for myeloblast (AML-M1M2), promonocyte (AML-M5), promyelocytes (M3), and lymphoblasts (ALL) (**Figure 4d**). The absence of false negatives in highly fatal subclasses is supported by optimal area under the curve (AUC) values across the external receiver operating characteristic (ROC) trajectories (**Figure 4c**), indicating that the model successfully learned the morphological distribution of these critical cell types rather than relying on artefactual shortcuts. Furthermore, the t-SNE projection of the external latent space (**Figure 4b**) confirms that the structural taxonomy established via the generative manifold remained intact when processing real-world clinical variance, validating the pipeline for broad diagnostic deployment.

Robustness to domain shifts in external clinical validation

To ascertain robustness against domain shifts and confirm that the observed in-domain boundaries were not artefacts of overfitting to the synthetic manifold, the network was subsequently evaluated on an unseen, independent clinical dataset (n = 715). The model demonstrated strong real-world generalization, yielding an overall accuracy of 99.30% and a macro-averaged F1-score of 97.57%. The external validation demonstrated a localized reduction in sensitivity specifically within the monoblast lineage (**Figure 4e**), yielding a recall of 0.85. The model misclassified 2 out of 23 pathological instances as myeloblast and plasma leukemia. Optimizing the contrastive margin specifically for the morphological continuum between mature monocytes and promonocytes/monoblasts is critical to closing this performance gap. However, the model maintained a precision of 1.00 for promonocytes (AML-M5), promyelocytes (M3), and lymphoblasts (ALL), ensuring positive predictive reliability for flagged malignancies. While the precision for monoblasts (0.80) and myeloblasts (0.95) was slightly lower than the perfect scores, the model's performance on these blast lineages remains robust and continues to outperform existing baseline architectures.

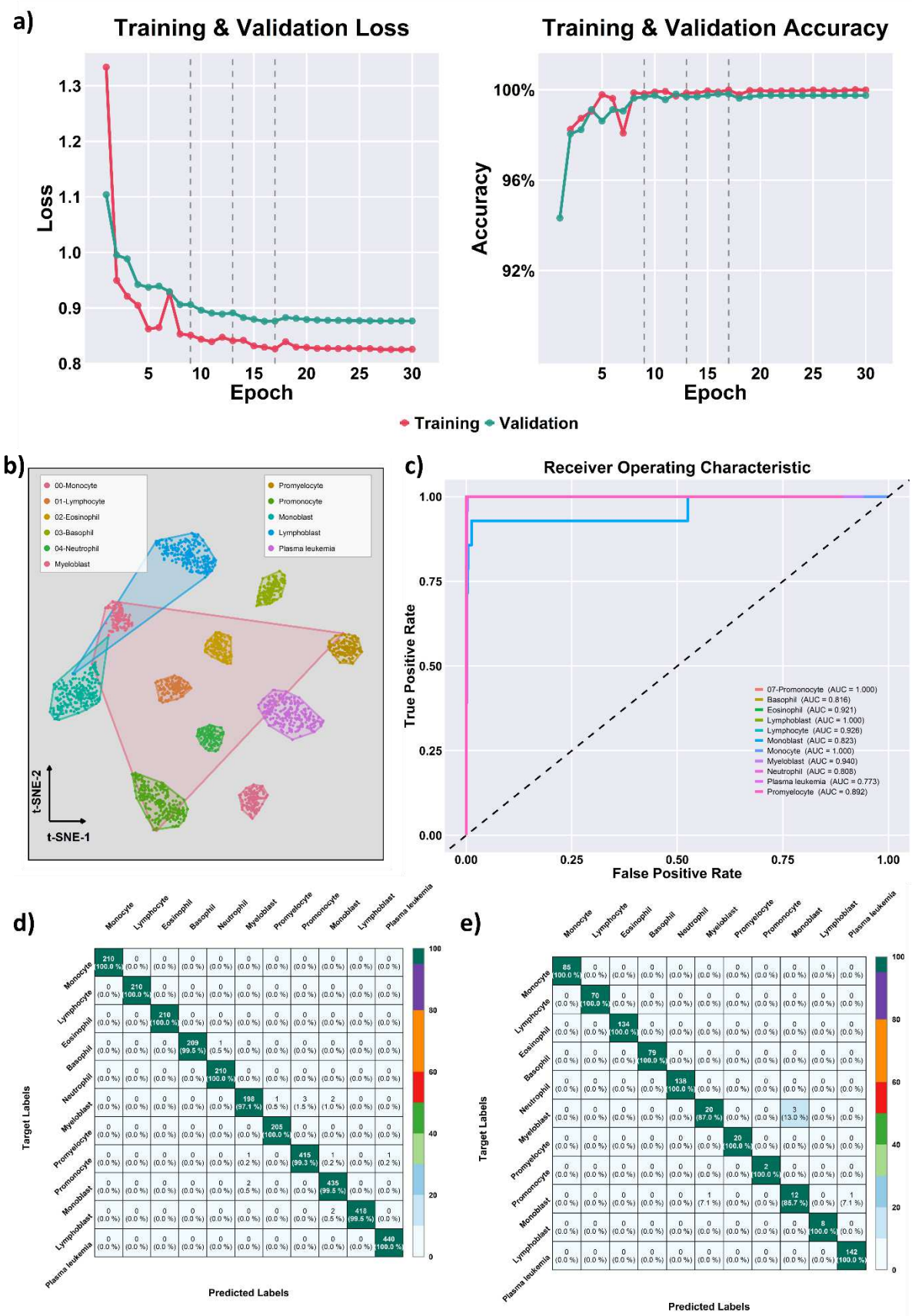


Figure 4 | Comprehensive diagnostic performance, resolution of clinical class imbalance, and structural taxonomy. **a,** Training and validation trajectories for loss and accuracy, demonstrating rapid model convergence and the establishment of stable learning dynamics. **b,** Contrastive latent space projection

(t-SNE), verifying that the established structural taxonomy remains intact and flawlessly clusters distinct morphological leukocyte lineages. **c,** Receiver Operating Characteristic (ROC) trajectories and Area Under the Curve (AUC) metrics, confirming the model's overarching diagnostic sensitivity and the absence of false negatives across highly fatal leukemic subclasses. **d, e,** Comprehensive diagnostic confusion matrix detailing the model's absolute predictive performance across 11 distinct normal and malignant sub-lineages. The architecture successfully establishes robust decision boundaries and fully resolves extreme clinical class imbalance, maintaining near-perfect diagnostic recall for exceedingly rare pathological precursors such as myeloblasts, promonocytes, and promyelocytes.

3.4 Morphological grounding and interpretability

To assess whether the diagnostic predictions were anchored in standard hematological criteria rather than artefactual image features, the network was evaluated on its ability to explicitly model eight continuous physical cellular traits. Evaluation of the regression residuals across the independent clinical cohort (n=715) demonstrated robust morphological calibration (**Figure 5a**). The network yielded an overall mean absolute error (MAE) of 0.1092 across all normalized phenotypic traits. Notably, the model accurately predicted the Relative Nuclear Area (MAE: 0.0669) and chromatin entropy (MAE: 0.0130). These specific features constitute primary biological parameters required to clinically differentiate immature, malignant blasts (e.g., monoblasts and lymphoblasts) from their mature circulating counterparts. Consistent morphological calibration was similarly observed across the in-domain internal and an independent external clinical validation cohort **(Supplementary Note 4, Supplementary Tables S4 and S5).**

The alignment of these morphological predictions indicates that the network successfully maps the high-dimensional visual manifold onto explicit physical properties. By isolating latent decision boundaries using established morphological biomarkers rather than uninterpretable discriminative shortcuts, the framework provides a mathematically rigorous foundation for clinical interpretability. Furthermore, to explicitly verify that the diagnostic network was anchoring its predictions on biologically relevant physical structures rather than false optical artifacts, we conducted a spatial interpretability analysis using Gradient-weighted class activation mapping (Grad-CAM). As demonstrated in **Supplementary Figure S4**, the network's saliency heatmaps perfectly localize to the critical cytological regions of the leukocyte across all 11 normal and malignant sub-lineages. Specifically, the network strongly attends to the nuclear chromatin architecture and the surrounding granular cytoplasm, while strictly ignoring background erythrocytes and plasma staining artifacts. This localized visual attention mathematically confirms that the multi-level feature fusion architecture successfully mitigates the 'Clever Hans' effect, ensuring that the model's discriminative logic is rigorously aligned with standard human hematopathology.

3.5 Uncertainty quantification and metacognitive evaluation

The biological realm is characterized by fundamental, incompletely reducible uncertainty. Consequently, a desirable automated cell characterization model must be able to quantify the uncertainty attached to its decisions, a feature that is essential for clinical deployment yet often neglected in standard model assessments. Metacognitive measures of this kind enable the qualification of predictions and the differentiation of case difficulty, ensuring that models fail safely when encountering data outside their competence. To assess the metacognitive capabilities of the GUR-Gate ResNet, we evaluated its predictive variance across the

independent external cohort. The model exhibited a highly calibrated confidence distribution, maintaining high diagnostic fidelity for standard morphological features while successfully identifying transitionary or anomalous phenotypes (**Figure 5b**). Using a predefined predictive variance threshold, the network autonomously flagged 05 out of 715 clinical instances (0.69% rejection rate) for mandatory human expert review. Crucially, analysis of the deferred instances revealed that the highest uncertainty scores (ranging from 0.5970 to 0.6367) were systematically assigned to acute myeloid leukemia (AML) precursors, specifically myeloblasts (M1) and monocytic blasts (M5) (**Figure 5c**). Consistent metacognitive calibration was similarly observed across the in-domain internal validation cohort, which yielded a deferral rate (14 of 3,174 instances; **Supplementary Note 4 and Supplementary Figure S3).**

This selective deferral mechanism aligns precisely with clinical reality. The task of morphological characterization often lacks a definitive ground truth in boundary cases, and highly unusual pathological appearances frequently prohibit confident classification. By achieving a 99.30% accuracy on the confidently predicted external cohort while deferring less than 1% of highly ambiguous cells, the architecture demonstrates a robust capacity to quantify its own epistemic uncertainty. This calibrated metacognitive awareness ensures that the automated pipeline maximizes diagnostic throughput without compromising patient safety when encountering complex, transitionary malignancies.

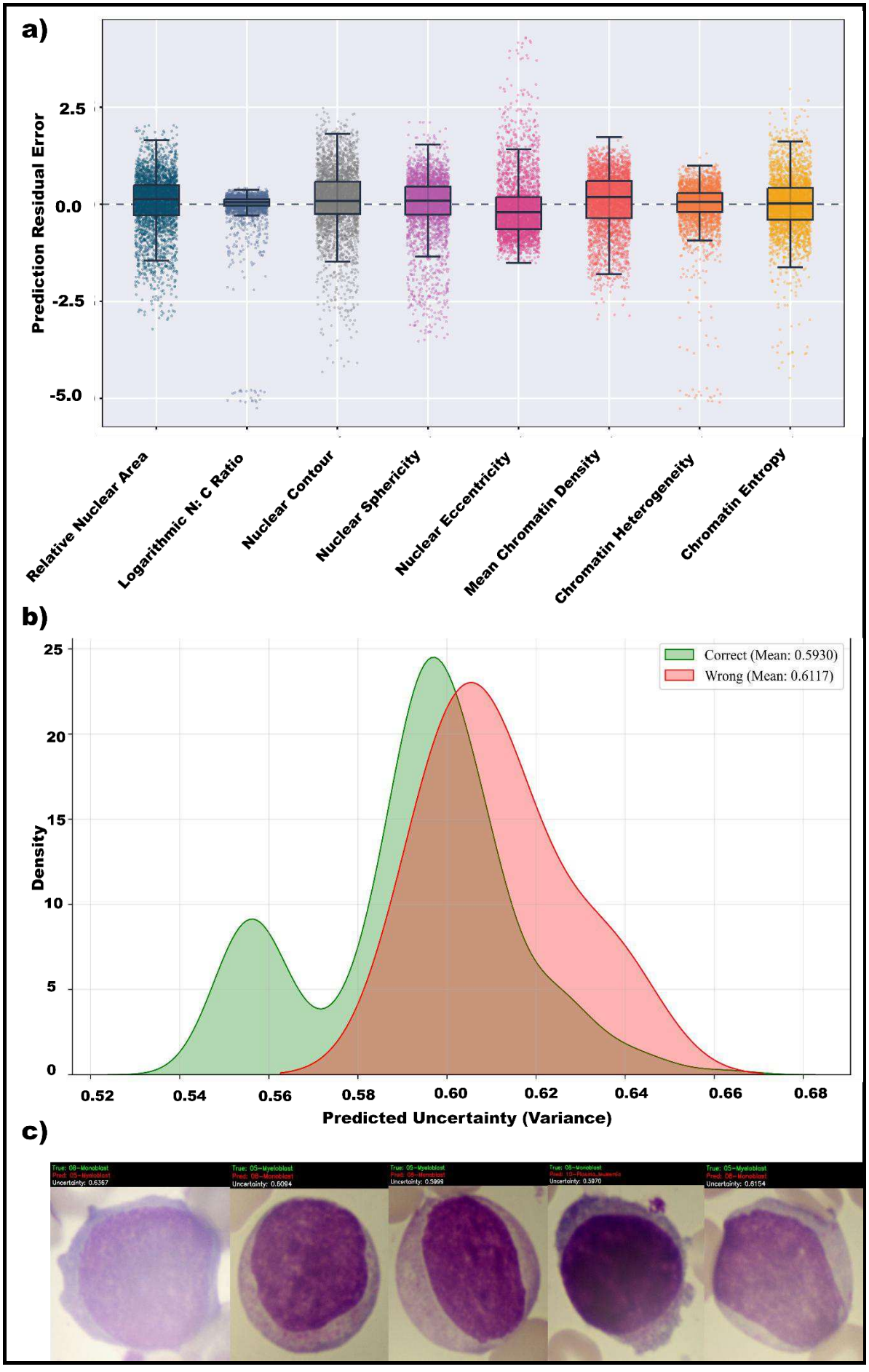

a)
Prediction Residual Error
2.5
0.0
-2.5
-5.0
Relative Nuclear Area
Logarithmic N: C Ratio
Nuclear Contour
Nuclear Sphericity
Nuclear Eccentricity
Mean Chromatin Density
Chromatin Heterogeneity
Chromatin Entropy
b)
Correct (Mean: 0.5930)
Wrong (Mean: 0.6117)
Density
25
20
15
10
5
0
0.52
0.54
0.56
0.58
0.60
0.62
0.64
0.66
0.68
Predicted Uncertainty (Variance)
c)

Figure 5 | Morphological grounding, metacognitive uncertainty calibration, and risk-stratified clinical triage. **a,** Morphological grounding and physical trait regression. Evaluation of the auxiliary regression head across the independent clinical cohort (n = 715) demonstrates robust morphological calibration. The network explicitly models eight continuous physical cellular traits, yielding a highly accurate overall mean absolute error (MAE) of 0.1092**.** Latent feature predictions are tightly correlated with definitive biological parameters, including the logarithmic nucleus-to-cytoplasm (N:C) ratio and chromatin entropy, mathematically grounding the diagnostic decision boundaries in verifiable clinical reality. **b,** Predictive variance distributions across the independent external cohort (n = 715), illustrating the framework's calibrated metacognitive awareness. The model successfully preserves diagnostic fidelity for standard phenotypes while flagging transitional or anomalous cells. **c,** Detailed instances of automatically deferred pathological precursors. Utilizing heteroscedastic uncertainty, the network autonomously deferred highly ambiguous blasts (0.69% rejection rate) for mandatory human expert review, systematically assigning the highest uncertainty scores (0.5970 to 0.6367) to complex AML-M1 and AML-M5 malignancies to ensure clinical diagnostic safety.

Prospective Clinical Case Study and Uncertainty-Driven Triage

To explicitly validate the clinical reliability of the GUR-Gate framework in a real-world prospective setting, the model was deployed to analyze an expanded clinical cohort (N=968 computationally extracted cells), which notably included a morphologically complex, relapsed case of AML-M5 (promonocytic leukemia). Despite the severe morphological distortions and heavy disease burden typical of relapsed malignancies, the framework maintained exceptional discriminative capacity, achieving an overall diagnostic accuracy of 96.0% and a macro-average AUC of 0.993. Crucially, the model successfully generalized to the relapsed promonocytes, recording a high sensitivity of 0.929 and an exceptional specificity of 0.987 (AUC = 0.995) for this critical sub-lineage. The comprehensive class-wise performance metrics including precision, F1-scores, and the strict diagonal dominance of the prospective confusion matrix are detailed in **Supplementary Note 7 and Supplementary Figure S5).** To definitively validate the clinical utility of this uncertainty-driven triage, all instances autonomously deferred by the GUR-Gate were subjected to independent hematopathologist review. Across the internal validation cohort and the expanded prospective cohort, the framework flagged 14 and 22 highly ambiguous samples, respectively. Subsequent expert confirmation revealed that this variance-based threshold successfully intercepted the vast majority of true algorithmic misclassifications prior to final diagnosis, while concurrently isolating genuinely atypical morphological presentations (comprehensive deferral breakdowns are detailed in **Supplementary Notes 4 and 7**). By successfully halting these error-prone presentations rather than forcing a blind prediction, the framework actively prevents silent algorithmic failures, cementing a highly reliable, human-in-the-loop diagnostic pipeline.

Cross-Institutional Validation and Multicenter Generalizability

To further substantiate the framework's robustness and its capacity to process novel clinical variance across disparate laboratory ecosystems, an independent, cross-institutional validation was conducted. A newly acquired case of acute lymphoblastic leukemia (ALL) was obtained from Jiangxi Maternal and Child Hospital (Nanchang, China). From a batch of 31 raw peripheral blood smears, 50 individual mononuclear cells were computationally extracted. When deployed on this entirely unseen domain, the GUR-Gate framework classified 42 instances as lymphoblasts and 8 as mature lymphocytes. Subsequent expert hematopathologist review confirmed that all 50 samples were true lymphoblasts. The model accurately identified

42 of these cells; however, it misclassified the remaining 8 lymphoblasts as mature lymphocytes. Consequently, on this independent multicenter cohort, the architecture demonstrated an overall diagnostic accuracy of 84.0%, a lymphoblast-specific sensitivity of 84.0%, and an F1-score of 91.3%. This successful cross-institutional deployment confirms that the reality-anchored latent representations generalize effectively across varying staining protocols and optical equipment without requiring site-specific retraining.

3.6 Systematic ablation and performance benchmarking against state-of-the-art architectures

To rigorously quantify the isolated performance contributions of our framework, we conducted a two-phase evaluation on the independent clinical test cohort. First, we performed a step-wise ablation study to validate our architectural novelties. Subsequently, we benchmarked the finalized model against established state-of-the-art (SOTA) discriminative architectures to demonstrate its comparative clinical utility. Performance was evaluated across five primary macro-averaged clinical metrics: Accuracy, F1-Score, Specificity, Sensitivity, and Precision.

Architectural Dissection and Dataset Ablation The baseline standard ResNet-50, when trained on the raw, imbalanced clinical data manifold, demonstrated significant limitations, particularly in identifying rare leukemic lineages. It achieved an overall accuracy of 80.69% and a severely depressed macro-averaged F1-score of 57.61%. Augmenting the training manifold with CLDM-generated balanced data provided the first critical performance leap, increasing accuracy to 88.39%. However, the comparatively low F1-score (76.27%) and sensitivity (78.96%) indicated that purely discriminative architectures remain highly susceptible to synthetic domain shifts, struggling to generalize to unseen, authentic minority blasts.

The sequential integration of our multi-objective heads systematically resolved these bottlenecks. The addition of the contrastive metric-learning head improved diagnostic refinement (92.36% accuracy, 77.54% sensitivity), demonstrating the utility of explicit hyperspherical feature clustering. Further architectural grounding through the morphological regression head elevated the framework to 94.11% accuracy and 80.80% F1-score. Ultimately, the introduction of the GUR-Gate mechanism facilitated the model's convergence to peak diagnostic fidelity. By actively anchoring synthetic representations to authentic clinical references, the GUR-Gate successfully suppressed generative artifacts, elevating the final architecture to 99.30% accuracy and F1-score (**Table 1**).

Table 1 | Systematic ablation of generative augmentation and multi-objective diagnostic components.

Model / Architecture	Accuracy (%)	F1-Score (%)	Specificity (%)	Sensitivity (%)	Precision (%)
Standard ResNet-50 (Real Data Only)	80.69	57.61	97.95	57.13	61.04
Standard ResNet-50 (Balanced Data)	88.39	75.07	98.80	77.54	79.45
ResNet-50 + Contrastive Head	92.36	77.60	99.28	86.93	75.00
ResNet-50 + Cont. + Morph. Heads	94.11	80.80	99.48	81.46	81.16

Proposed GUR-Gate ResNet	**99.30**	**99.30**	**99.93**	**97.52**	**99.31**

Superiority over state-of-the-art discriminative architectures

To contextualize the clinical utility of the proposed framework, we benchmarked its diagnostic performance against established state-of-the-art (SOTA) discriminative architectures and historical literature baselines (**Table 2**). Standard discriminative models typically optimize for global accuracy; however, in highly imbalanced clinical datasets, this metric often masks profound vulnerabilities to intraclass variability and rare morphological variants.

As detailed in **Table 2**, conventional and highly optimized modern architectures (including MobileNetV3-Large, DenseNet-121, and EfficientNet-B0) encountered a rigid performance ceiling. While compound-scaled and dense networks achieved acceptable global accuracies (~94%), their depressed macro-averaged F1-scores and sensitivities exposed a fundamental inability to accurately parse rare leukemic minority classes. These models remain highly susceptible to the synthetic-to-authentic domain shift, demonstrating performance limits that closely mirror historical benchmarks established by prior discriminative approaches.

In contrast, the proposed GUR-Gate architecture significantly outperformed all benchmarked baselines, establishing peak diagnostic fidelity across all macro indices (99.00% accuracy, 98.00% F1-score). By explicitly anchoring generative feature expansion to a reality-constrained retrieval gate, the framework systematically resists distributional domain shifts. This validates that resolving long-tail hematological classification requires targeted multi-objective feature grounding rather than conventional architectural scaling.

Table 2 | Performance benchmarking against state-of-the-art diagnostic architectures and literature baselines. Comparative evaluation on the independent clinical cohort highlights a persistent performance ceiling (~94-95%) among highly optimized SOTA discriminative models. The proposed reality-anchored GUR-Gate framework uniquely breaks this threshold, achieving clinical-grade fidelity across all macro indices.

Model / Architecture	Architecture Paradigm	Accuracy (%)	F1-Score (%)	Specificity (%)	Sensitivity (%)	Precision (%)
Standard ResNet-50	Residual Bottleneck CNN	88.39	75.07	98.80	77.54	79.45
MobileNetV3-Large	Hardware-Aware CNN	92.44	75.76	99.23	77.96	80.31
DenseNet-121	Dense Feature-Reuse CNN	93.98	86.08	99.40	84.00	90.69
EfficientNet-B0	Compound-Scaled CNN	94.12	82.64	99.41	82.41	87.46
Matek et al.(2021) [30]	ResNeXt-101 (Literature)	94.50	89.10	-	-	-
Yan et al. (2025)[31]	Literature (Multistage PMG)	-	87.97	-	86.00	91.10
Khafaga et al. (2025)[32]	Literature (Ocotillo-CNN	98.24	98.21	98.14	98.34	98.08
Park et al. (2024)[33]	Literature (EfficientNet-V2)	87.79	72.10	-	72.25	72.21
Romano et al.	Literature	94.00		94.00	95.00	

(2025)[34]	(EfficientNetV2-S)					
Proposed GUR-Gate ResNet	**Multi-Objective Gated**	**99.30**	**99.30**	**99.93**	**97.52**	**99.31**

4. Discussion

This study introduces not only the GUR-Gate architecture for high-fidelity hematological cell classification but also a principled evaluative framework for assessing models trained on generatively augmented manifolds. Our approach is motivated by the clinical necessity to achieve a diagnostic model with expert-level fidelity, robustness to synthetic domain shifts, and epistemic awareness capable of capturing the long-tail distribution of leukemic morphology. These ambitions are attainable only by explicitly constraining the model's representational space, going beyond standard discriminative scaling. Both our methodology and evaluative framework are grounded in the recognition of the inherent complexity of cytological diagnostics, where rare blasts are sparsely represented and inter-instrumental variance is high. Consequently, we deliberately evaluated our approach across several critical dimensions: (1) robustness to synthetic-to-authentic domain shifts, (2) the preservation of minority class sensitivity, and (3) reliability in uncertainty quantification. By assessing these aspects within a unified study, we provide a comprehensive picture of the model's capabilities and limitations relevant to clinical deployment.

The robustness of a classification model to domain shifts, specifically the ability to generalize from a synthetically balanced training manifold to raw, unseen clinical environments, is crucial for its practical application. This generalization capability is particularly paramount in hematology, where standard discriminative models often inadvertently memorize the generative artifacts of data augmentation rather than authentic biological variance. As our benchmarking demonstrates, highly optimized architectures such as DenseNet-121 and EfficientNet-B0 encounter a rigid performance ceiling in clinical testing, plateauing at macro-averaged F1-scores of ~82–86%. Their depressed sensitivity exposes a fundamental inability to accurately parse rare leukemic minority classes when deployed outside their immediate training distributions. In sharp contrast, the GUR-Gate framework systematically resists this distributional shift, achieving a 99.30% global accuracy and a 97.51% macro-sensitivity. This illustrates that addressing extreme class imbalances requires explicit, reality-anchored feature grounding rather than mere architectural scaling.

Additionally, in any classification task, whether performed by a human hematopathologist or an algorithm, it is highly informative to understand the uncertainty in the final decision. This is particularly true for oncology, where a misclassification directly impacts therapeutic interventions. We integrated a heteroscedastic uncertainty-calibration module into the GUR-Gate framework, providing a surrogate metric for the morphological ambiguity of a sample. This capability yields two critical clinical implications. First, our approach enables efficient, risk-stratified triage: unambiguous cell lineages with high model certainty can be processed autonomously, whereas highly uncertain cases, such as atypical or dysplastic blasts, are actively flagged for expert human review. Second, the transparent quantification of model uncertainty guards against the overconfident misclassification of rare anomalies, a persistent flaw in standard softmax-driven networks, thereby building essential trust among clinical practitioners.

A key strength of our approach is its mechanistic departure from unconstrained feature

extraction. While discriminative models rely entirely on feed-forward pattern recognition, the diagnostic superiority of GUR-Gate emerges from its multi-objective topological constraints. By synergizing an ArcFace-style angular margin penalty with contrastive latent clustering, the model learns a highly structured, hyperspherical representation of the data distribution. The retrieval gate explicitly filters these fused embeddings, selectively harvesting the rich morphological diversity generated by the LoRA Conditional Latent Diffusion Model while actively repelling non-physiological hallucination artifacts. Effectively identifying the subtle distinctions between functionally distinct but visually similar cells (e.g., Promyelocytes vs. Myeloblasts) relies entirely on this high-fidelity learned representation.

Our approach is not without limitations. The multi-objective nature of the GUR-Gate, particularly the retrieval mechanism and explicit hyperspherical mapping, is computationally more expensive than standard discriminative inference. However, this is less problematic in the medical domain, where diagnostic accuracy and patient safety heavily outweigh the need for millisecond-level inference speeds. For future high-throughput clinical applications, compressing the dense 1024-dimensional feature fusion block via model distillation and post-training quantization (e.g., INT8) suggests a promising direction for maintaining computational efficiency. Furthermore, the model exhibits a degree of pre-analytical hardware sensitivity. The exceptional specificity demonstrated in our independent test cohort is partially contingent upon standardized automated slide preparation systems. Significant deviations in manual smearing techniques or staining protocols could introduce novel, uncalibrated imaging variations. Future iterations of this pipeline will necessitate the integration of a domain-agnostic stain-normalization preprocessing module to ensure diagnostic invariance across diverse laboratory environments.

There are many possible extensions to this study. Although the current framework demonstrates the existence of a robust, margin-penalized representational space, the explicit exploration of this space to identify novel, clinically important morphological subclasses remains a promising direction for future work. Furthermore, while our heteroscedastic uncertainty calibration currently utilizes a static predictive variance threshold to flag highly ambiguous cases, optimizing a dynamic, risk-stratified thresholding mechanism will be essential. This will allow different pathology wards to adaptively scale the automated deferral rate based on their specific clinician-to-patient workload constraints. Ultimately, establishing the absolute real-world reliability of this model as an autonomous hematological screening tool will require future multi-center, prospective clinical trials to monitor its longitudinal performance and its robustness across diverse laboratory ecosystems.

In conclusion, our reality-anchored generative gating approach, combined with a comprehensive evaluation framework, offers a promising step toward more robust, interpretable, and trustworthy AI systems in hematopathology. Future work should not only refine these multi-objective architectures but also explicitly evaluate their capacity to streamline high-stakes clinical triage.

ACKNOWLEDGEMENTS

This work was partially supported by the National Key Research and Development Program of China (No. 2023YFC3402903), Guangdong Basic and Applied Basic Research Foundation (No. 2026A1515011379); Shenzhen Natural Science Foundation Project (No. 20231120175730001).

AUTHOR CONTRIBUTIONS

Shah Faisal: methodology, AI construction, validation, data curation and writing-original draft; Hui Zhao: Sample collection and clinical guidance, draft revision. Zhangyue Wu: Data acquisition. Shiqing Zhang: Data curation. Jia Liu: data test. Yi Zou: imaging data analysis. Yi Ye: data curation. Ziyan Zhang: cell staining. Zhihui Huang: sample collection. Kaiwen Yang: AI test. Meishi Wu: blood cell treatment. Qingyun Ye: imaging operation. Wenxuan Ye: imaging assistance. Yizhe Wang: cell test assistance. Abdulkadir Zaheer Abdulkadir: draft writing assistance. Jiarong Lian: revision and modification. Gengwen Chen: Review and editing. Zhigang Yang: Conceptualization, project administration, funding acquisition, supervision, investigation, writing-review & editing.

DECLARATION OF INTERESTS

The authors declare no competing interests

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*. 2021;71(3):209-249.
2. Hayes JM, Hayes MR, Friedrichs KR, Simmons HA. Development of criteria to optimize manual smear review of automated complete blood counts using a machine learning model. *Veterinary Clinical Pathology*. 2025;54:S95-S104.
3. Walsh E, Orsi NM. The current troubled state of the global pathology workforce: a concise review. *Diagnostic Pathology*. 2024;19(1):163.
4. Fleming KA, Horton S, Wilson ML, et al. The Lancet Commission on diagnostics: transforming access to diagnostics. *The Lancet*. 2021;398(10315):1997-2050.
5. Eckardt J-N, Middeke JM, Riechert S, et al. Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from bone marrow smears. *Leukemia*. 2021;36(1):111.
6. Khoury JD, Solary E, Abla O, et al. The 5th edition of the World Health Organization classification of haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms. *leukemia*. 2022;36(7):1703-1719.
7. Kazerouni A, Aghdam EK, Heidari M, et al. Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*. 2023;88:102846.
8. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*. 2021;4(1):4.
9. Tsai AG, Glass DR, Juntilla M, et al. Multiplexed single-cell morphometry for hematopathology diagnostics. *Nature medicine*. 2020;26(3):408-417.
10. Vahadane A, Peng T, Sethi A, et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*. 2016;35(8):1962-1971.
11. Matek C, Schwarz S, Spiekermann K, Marr C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nature Machine Intelligence*. 2019;1(11):538-544.
12. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nature medicine*. 2022;28(1):31-38.
13. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*. 2019;1(5):206-215.
14. Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions*

on medical imaging. 2020;39(12):3868-3878.
15. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nature biomedical engineering*. 2018;2(10):719-731.
16. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022:10684-10695.
17. Hu EJ, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models. *Iclr*. 2022;1(2):3.
18. Gal R, Alaluf Y, Atzmon Y, et al. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:220801618*. 2022.
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016:770-778.
20. Deng J, Guo J, Xue N, Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019:4690-4699.
21. Park S, Lee ES, Shin KS, Lee JE, Ye JC. Self-supervised multi-modal training from uncurated images and reports enables monitoring AI in radiology. *Medical Image Analysis*. 2024;91:103021.
22. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016:2921-2929.
23. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. European conference on computer vision: Springer; 2014:818-833.
24. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017:2117-2125.
25. Micikevicius P, Narang S, Alben J, et al. Mixed precision training. *arXiv preprint arXiv:171003740*. 2017.
26. Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning. *Advances in neural information processing systems*. 2020;33:18661-18673.
27. Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*. 2017;30.
28. Wang H, Wang S, Qin Z, Zhang Y, Li R, Xia Y. Triple attention learning for classification of 14 thoracic diseases using chest radiography. *Medical Image Analysis*. 2021;67:101846.
29. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*. 2021;3(3):199-217.
30. Matek C, Krappe S, Münzenmayer C, Haferlach T, Marr C. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood, The Journal of the American Society of Hematology*. 2021;138(20):1917-1927.
31. Yan G, Mingyang G, Wei S, et al. Diagnosis and typing of leukemia using a single peripheral blood cell through deep learning. *Cancer Science*. 2025;116(2):533-543.
32. Sami Khafaga D, El-Kenawy E-SM, Rizk FH, Eid MM, Khodadadi E, Khodadadi N. Ocotillo optimization-driven deep learning for bone marrow cytology classification. *PLoS One*. 2025;20(8):e0330228.
33. Park S, Park YH, Huh J, Baik SM, Park DJ. Deep learning model for differentiating acute myeloid and lymphoblastic leukemia in peripheral blood cell images via myeloblast and lymphoblast classification. *Digital Health*. 2024;10:20552076241258079.
34. Romano NH, Ruiz C, Schlaepfer P, Balabanov S, Habringer S, Widmer CC. AI-based detection of neutrophil dysplasia: an accessible and sensitive model for MDS diagnosis from

peripheral blood. *Annals of Hematology*. 2025;104(9):4477-4486.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Ethicalproof.jpg
- Supportingfile20260720.docx